John von Neumann Institute for Computing

Evolutionary Analysis in Pathogenesis-Related Proteins

Nicole M. Scherer, Claudia E. Thompson,
Loreta B. Freitas, Sandro L. Bonatto,
Francisco M. Salzano

published in

NIC Workshop 2006,
From Computational Biophysics to Systems Biology,
Jan Meinke, Olav Zimmermann,
Sandipan Mohanty, Ulrich H.E. Hansmann (Editors)
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **34**, ISBN-10: 3-9810843-0-6,
ISBN-13: 978-3-9810843-0-6, pp. 193-196 , 2006.

http://www.fz-juelich.de/nic-series/volume34

Evolutionary Analysis in Pathogenesis-Related Proteins

Nicole M. Scherer[1], Claudia E. Thompson[2], Loreta B. Freitas[2], Sandro L. Bonatto[3], and Francisco M. Salzano[2]

[1] Bioinformatics Department, Heinrich-Heine-Universität Düsseldorf
Universitätsstrasse 1, Geb. 25.02.02, 40225 Düsseldorf, Germany
E-mail: scherer@cs.uni-duesseldorf.de

[2] Departamento de Genética, Instituto de Biociências, UFRGS
Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brazil
E-mail: francisco.salzano@ufrgs.br

[3] Centro de Biologia Genômica e Molecular, Faculdade de Biociências, PUCRS
Av. Ipiranga 6681, 90619-900 Porto Alegre, RS, Brazil

Pathogenesis-related proteins (PRs) are expressed by host plants following infections by fungi, bacteria or viruses, or after induction by abiotic stress factors. PRs enhance the host capability to limit subsequent infections. They comprise a wide range of forms, such as hydrolases, transcription factors, protease inhibitors, enzymes associated with various metabolic pathways, and allergenic products. Their functional motifs are related to a number of eukaryotic proteins, involved in very distinct functions. It is possible that their defensive functions evolved after their emergence as gene families. We believe that natural selection is a fundamental factor in the establishment of this resistance form in plants. Our analysis on the primary structure of representatives of 14 PR families identified several target sites for adaptive evolution. Testing how these changes structurally affect the protein molecules should help to relate adaptive mutations with their biological functions.

1 Introduction

Plants are likely to be infected by a large number of pathogens, like bacteria, fungi and viruses. Hence, plants have evolved a variety of mechanisms to prevent pathogen colonization and disease. Our study deals with a special class of defense proteins called pathogenesis-related proteins (PRs). Until now, 17 PR families have been identified[1]. They are induced after pathogenic infection or environmental stress, and exhibit antifungal, antibacterial, insecticidal, nematicidal and antiviral effects. These provide the host plant enhanced capability to limit subsequent infections inhibiting pathogen growth, multiplication or spread. Their toxicity is due to hydrolicity, proteinase-inhibitory and membrane-permeabilizing ability[2].

In 2000, Bishop *et al.*[3] combined sequence evolution analysis and knowledge of the structure of the chitinase type I gene family to understand the variable effectiveness of specific chitinases against different pathogens. They found an excess of amino acid replacements in the active site and substrate binding cleft, which cannot be explained by a relaxation of selection pressure alone and therefore indicates positive selection.

In this study, we first conduct a phylogenetic analysis of pathogenesis-related proteins with protein-coding DNA sequences of 14 PR families[4]. We then apply a maximum likelihood framework based on codon substitution models to identify adaptive evolution. Finally, the sites identified as positively selected are further investigated on the structure of the protein.

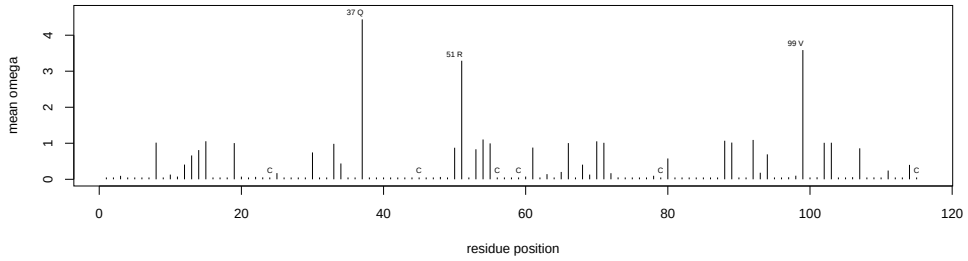

Figure 1. Values for ω ratios for sites along the PR-4 sequence under model M2 (Positive Selection) in CODEML[8]. Cysteine residues involved in disulphide bridges are indicated by the letter C

2 Methods

Not all mutations occurring in the DNA sequences lead to amino acid substitutions in the protein sequence. Those mutations which do change the coded amino acid are called non-synonymous. Synonymous mutations, in contrast, are silent and do not alter the amino acid sequence. Thus, they are considered free from selection pressure. Positive selection at the molecular level is normally tested by the comparison of non-synonymous (d_N) and synonymous (d_S) substitution rates in protein-coding genes. The ratio $\omega = d_N/d_S$ is an indicator for selective pressure. Positive selection is indicated by $\omega > 1$. On the other hand, $\omega < 1$ indicates purifying selection and $\omega = 1$ neutral evolution.

For each PR family, typical patterns and the reference sequences were used to search for homologous sequences in the Swiss-Prot/TrEMBL (UniProt Knowledgebase[5]) protein sequence database. We used ClustalX[6] to align the sequences of each dataset. The protein and DNA sequence alignments were compared and manually edited with respect to the codon frame. Then, we reconstructed the phylogenetic trees of each PR family using IQPNNI[7]. Based on these trees, we applied CODEML[8] to obtain the values of ω for each site under different models. Nested models like M1 (nearly neutral) and M2 (positive selection) are compared in a likelihood ratio test (LRT). If M1 is rejected in favor of M2 we assume that the sites with $\omega > 1$ are under positive selection. We applied SWISS-MODEL[9] to infer the protein structure of all our PR dataset, except for PR-7, for which no appropriate template was available. Finally, we calculated the Euclidean distance from the alpha-carbon of each amino acid to the position of its homologous site in the template.

3 Results

The maximum likelihood analysis of the 14 PR datasets indicates positive selection in eight families of pathogenesis-related proteins. Here, we present the results for PR-4 (a wound-induced chitin-binding protein) as an example. Figure 1 shows the ω values for each residue position obtained under model M2 in CODEML[8]. The positively selected sites (37 Q, 51 R, 99 V) are labeled. Residues 37 Q and 99 V are placed in alpha-helices, and 51 R belongs to a gamma-turn. The members of this family have six conserved cysteine residues that form three disulphide bridges[10]. The six cysteine residues are among the sites with ω ratios close to zero due to purifying selection.

4 Discussion

The Euclidean distances calculated on the superimposed structures should provide a means to examine the relationship between the positively selected sites and their function on the protein. In the inferred structures of the PRs, the differences observed in the positively selected sites are due to differences in residue size. The deviations resulting from *indels* are greater than those resulting from substitutions. Using only this approach we could not see a correlation between these mutations and their effects in the protein structure. In order to make inferences about the positively selected sites, it is necessary to include the biochemical and biophysical variables in the analysis, and also to treat the *indels* separately. In the future, other modeling approaches should be used to overcome these difficulties.

Acknowledgments

We would like to thank Roland Fleißner, Simone Linz and Bui Quang Minh for their helpful comments and suggestions on the poster and this manuscript. This work was supported by grants from PRONEX, CNPq, FINEP, CAPES, FAPERGS and PROPESQ-UFRGS.

References

1. `http://www.bio.uu.nl/~fytopath/PR-families.htm`
2. Aglika Edreva. *Pathogenesis-related proteins: research progress in the last 15 years*. Gen. Appl. Plant Physiology **31**, 105–124 (2005)
3. J. G. Bishop, A. M. Dean and T. Mitchell-Olds. *Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution*, PNAS **97**, 5322–5327 (2000)
4. N. M. Scherer, C. E. Thompson, L. B. Freitas, S. L. Bonatto, and F. M. Salzano. *Patterns of molecular evolution in pathogenesis-related proteins*, Genet. Mol. Biol. **28**, 645–653 (2005).
5. A. Bairoch , R. Apweiler, C.H. Wu, W.C. Barker, B Boeckmann., S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, L. S. Yeh. *The Universal Protein Resource (UniProt)*, Nucleic Acids Res. **33**, D154–D159 (2005) `http://www.uniprot.org`
6. R. Chenna, H. Sugawara, T. Koike, R. Lopez,T. J. Gibson, D. G. Higgins, J. D. Thompson. *Multiple sequence alignment with the Clustal series of programs.* Nucleic Acids Res. **31**, 3497–3500 (2003)
7. L. S. Vinh and A. von Haeseler. *IQPNNI: Moving fast through tree space and stopping in time*, Mol. Biol. Evol. **21**, 1565–1571 (2004)
8. Z. Yang. *PAML: a program package for phylogenetic analysis by maximum likelihood*, Comput. Appl. Biosci. **13**, 555–556 (1997)
9. T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch. *SWISS-MODEL: an automated protein homology-modeling server*, Nucleic Acids Research **31**, 3381–3385 (2003).
10. B. Svensson, I. Svendsen, P. Hojrup, P. Roepstorff, S. Ludvigsen and F. M. Poulsen. *Primary structure of barwin: a barley seed protein closely related to the C-terminal domain of proteins encoded by wound-induced plant genes.* Biochemistry **31**, 8767–8770 (1992)